CONFIDENCE INTERVALS FOR LOW-DIMENSIONAL PARAMETERS WITH HIGH-DIMENSIONAL DATA

CUN-HUI ZHANG AND STEPHANIE S. ZHANG

Abstract. The purpose of this paper is to propose methodologies for statistical inference of low-dimensional parameters with high-dimensional data. We focus on constructing confidence intervals for individual coefficients and linear combinations of several of them in a linear regression model, although our ideas are applicable in a much broad context. The theoretical results presented here provide sufficient conditions for the asymptotic normality of the proposed estimators along with a consistent estimator for their finite-dimensional covariance matrices. These sufficient conditions allow the number of variables to far exceed the sample size. The simulation results presented here demonstrate the accuracy of the coverage probability of the proposed confidence intervals, strongly supporting the theoretical results.

Key words: Confidence interval, p-value, statistical inference, regression, high dimension.

1. Introduction

High-dimensional data is an intense area of research in statistics and machine learning, due to the rapid development of information technologies and their applications in scientific experiments and everyday life. Enormous amounts of large complex datasets have been collected and are waiting to be analyzed; meanwhile, an enormous effort has been mounted in order to meet this challenge by researchers and practitioners in statistics, computer science, and other disciplines. A great number of statistical methods, algorithms, and theories have been developed for the prediction and classification of future outcomes, the estimation of high-dimensional objects for different purposes, and the selection of important variables or features for further scientific experiments and engineering applications. However, statistical inference with high-dimensional data is still a largely untouched territory due to the complexity of the sampling distributions of existing estimators. This is particularly the case in the context of the so called large-p-smaller-n problem, where the dimension of the data is greater than the sample size,

Linear regression is one of the best understood statistical models in high-dimensional data. Important work has been done in problem formulation, development, and analysis of methodologies and algorithms, and theoretical understanding of their performance under sparsity assumptions. This includes ℓ_1 regularized methods and their analysis [Tib96, CDS01, CT07, GR04, Gre06, MB06, Tro06, ZY06, Wai09, CT07, ZH08, MY09, BRT09, Kol09, Zha09, vdGB09, YZ10, KLT11, SZ11], nonconvex penalized methods [FF93, FL01, FP04, ZL08, KCO08, Zha10, ZZ11], greedy methods [Zha11a], adaptive methods [Zou06, HMZ08, Zha11b, ZZ11], screening methods [FL08], and more. For further discussion, we refer to related sections in the recent book [BvdG11] and recent reviews in [FL10, ZZ11].

Among existing results, variable selection consistency is most relevant to statistical inference. An estimator is variable selection consistent if it selects the oracle model composed of exactly the set of variables with nonzero regression coefficients. In the large-p-smaller-n setting, variable selection consistency has been established under incoherence and other ℓ_∞-type conditions on the design matrix for the Lasso [MB06, Tro06, ZY06, Wai09], and under sparse eigenvalue or ℓ_2-type conditions for nonconvex methods [FP04, Zha11a, Zha10, Zha11b, ZZ11]. Another approach in variable selection with high-dimensional data is subsampling or randomization methods, notably the stability selection method proposed in [MB10]. Since the oracle model is typically and nearly necessarily assumed to be of smaller order in dimension than the sample size n in selection consistency theory, consistent variable selection allows a great reduction of the complexity of the analysis from a large-p-smaller-n problem to one involving the oracle set of variables only. Consequently, taking the least squares estimator on the selected set of variables if necessary, statistical inference can be justified in the smaller oracle model. However, statistical inference based on selection consistency theory typically requires that all nonzero regression coefficients be greater than a noise level inflated to take model uncertainly into account. This assumption of either none or uniformly strong signal strength for all individual variables is, unfortunately, seldom supported by either the data or the underlying science, especially in biological and medical applications.

2. Methodology

We develop methodologies and algorithms for the construction of confidence intervals for the individual regression coefficients and their linear combinations in the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}), \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is a response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n\times p}$ is a design matrix with columns $\boldsymbol{x}_j$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown regression coefficients. The design matrix $\boldsymbol{X}$ is assumed to be deterministic throughout the paper, except in Subsection 3.3.

The following notation will be used. For vectors $\boldsymbol{v} = (v_1, \ldots, v_m)$ of any dimension, $\text{supp}(\boldsymbol{v}) = \{j : v_j \neq 0\}$, $|\boldsymbol{v}|_0 = |\text{supp}(\boldsymbol{v})| = \#\{j : v_j \neq 0\}$, and $|\boldsymbol{v}|_q = \{\sum_j |v_j|^q\}^{1/q}$, with the usual extension to $q = \infty$. For $A \subset \{1, \ldots, p\}$, $\boldsymbol{v}_A = (v_j, j \in A)^T$ and $\boldsymbol{X}_A = (\boldsymbol{x}_k, k \in A)$, including $A = -j = \{1, \ldots, p\} \setminus \{j\}$.

2.1. Bias corrected linear estimators.

In the classical theory of linear models, the least squares estimator of an estimable regression coefficient β_j can be written as

$$\widehat{\beta}_j^{(lse)} := \boldsymbol{y}^\top \boldsymbol{x}_j^\perp / |\boldsymbol{x}_j^\perp|_2^2, \tag{2}$$

where $\boldsymbol{x}_j^\perp$ is the projection of $\boldsymbol{x}_j$ to the orthogonal complement of the column space of $\boldsymbol{X}_{-j} = (\boldsymbol{x}_k, k \neq j)$. For estimable β_j and β_k,

$$\text{Cov}(\widehat{\beta}_j^{(lse)}, \widehat{\beta}_k^{(lse)}) = \sigma^2 (\boldsymbol{x}_j^\perp)^T \boldsymbol{x}_k^\perp / (|\boldsymbol{x}_j^\perp|_2\, |\boldsymbol{x}_k^\perp|_2). \tag{3}$$

In the high-dimensional case $p > n$, $\text{rank}(\boldsymbol{X}_{-j}) = n$ for all j when $\boldsymbol{X}$ is in general position. Consequently, $\boldsymbol{x}_j^\perp = 0$ and (2) is undefined. However, it may still be interesting to preserve certain properties of the least squares estimator. One advantage of (2) is the explicit formula (3) of the covariance structure. This feature holds for all linear estimators of $\boldsymbol{\beta}$. For any score vector $\boldsymbol{z}_j$ not orthogonal to $\boldsymbol{x}_j$, the corresponding linear estimator satisfies

$$\widehat{\beta}_j^{(lin)} = \frac{\boldsymbol{z}_j^T \boldsymbol{y}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} = \beta_j + \frac{\boldsymbol{z}_j^T \boldsymbol{\varepsilon}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} + \sum_{k \neq j} \frac{\boldsymbol{z}_j^T \boldsymbol{x}_k \beta_k}{\boldsymbol{z}_j^T \boldsymbol{x}_j}$$

with a similar covariance structure to (3). A problem with such a linear estimator is its bias. For every $k \neq j$ with $\boldsymbol{z}_j^T \boldsymbol{x}_k \neq 0$, the contribution of β_k to the bias is linear in β_k. In the worst case scenario where $\mathrm{sgn}(\beta_k) = \mathrm{sgn}(\boldsymbol{z}_j^T \boldsymbol{x}_k)$ for all k, the bias of $\widehat{\beta}_j^{(lin)}$ may exceed the order of existing ℓ_1 error bounds for the estimation of the entire vector $\boldsymbol{\beta}$. We note that for $\mathrm{rank}(\boldsymbol{X}_{-j}) = n$, it is impossible to have $\boldsymbol{z}_j \neq 0$ and $\boldsymbol{z}_j^T \boldsymbol{x}_k = 0$ for all $k \neq j$, so that bias is unavoidable. Still, this simple inspection of the bias of the linear estimator suggests a bias correction with an initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$:

$$\widehat{\beta}_j = \widehat{\beta}_j^{(lin)} - \sum_{k \neq j} \frac{\boldsymbol{z}_j^T \boldsymbol{x}_k \widehat{\beta}_k^{(init)}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} = \frac{\boldsymbol{z}_j^T \boldsymbol{y}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} - \sum_{k \neq j} \frac{\boldsymbol{z}_j^T \boldsymbol{x}_k \widehat{\beta}_k^{(init)}}{\boldsymbol{z}_j^T \boldsymbol{x}_j}, \tag{4}$$

with a score vector $\boldsymbol{z}_j$ depending on $\boldsymbol{X}$ only. One may also interpret (4) as a one-step self bias correction from the initial estimator,

$$\widehat{\beta}_j := \widehat{\beta}_j^{(init)} + \frac{\boldsymbol{z}_j^T \{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}\}}{\boldsymbol{z}_j^T \boldsymbol{x}_j}.$$

The estimation error of (4) can be decomposed as a sum of noise and approximation error:

$$\widehat{\beta}_j - \beta_j = \frac{\boldsymbol{z}_j^T \boldsymbol{\varepsilon}}{\boldsymbol{z}_j^T \boldsymbol{x}_j} + \frac{1}{\boldsymbol{z}_j^T \boldsymbol{x}_j} \sum_{k \neq j} \boldsymbol{z}_j^T \boldsymbol{x}_k (\beta_k - \widehat{\beta}_k^{(init)}). \tag{5}$$

A full description of (4) still require specifications of the score vector $\boldsymbol{z}_j$ and the initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$. These choices will be discussed in the following two subsections.

2.2. **Low-dimensional projections.** A proper choice of $\boldsymbol{z}_j$ should control both the noise and approximation error terms in (5), given suitable conditions on $\{\boldsymbol{X}, \boldsymbol{\beta}\}$ and an initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$. Recall that $\boldsymbol{z}_j$ aims to play the role of $\boldsymbol{x}_j^\perp$, the projection of $\boldsymbol{x}_j$ to the orthogonal complement of the column space of $\boldsymbol{X}_{-j} = (\boldsymbol{x}_k, k \neq j)$. When $|\boldsymbol{x}_j^\perp|_2$ is not too small, we may simply take $\boldsymbol{z}_j = \boldsymbol{x}_j^\perp$. When $|\boldsymbol{x}_j^\perp|_2$ is too small, e.g. $\boldsymbol{x}_j^\perp = 0$ when $\mathrm{rank}(\boldsymbol{X}_{-j}) = n$, we may use a $\boldsymbol{z}_j$ proportional to a relaxed projection of $\boldsymbol{x}_j$. Since $\boldsymbol{z}_j$ is a relaxed projection of $\boldsymbol{x}_j$ and the estimator (4) is given by a bias-corrected projection of $\boldsymbol{y}$ to the direction of $\boldsymbol{z}_j$, hereafter we will call (4) the low-dimensional projection estimator (LDPE) for easy reference.

The projection $\boldsymbol{x}_j^\perp$ is the residual of the least squares fit of $\boldsymbol{x}_j$ on $\boldsymbol{X}_{-j}$. A familiar relaxation of the least squares method is to add an ℓ_1 penalty. This leads to the choice of

$\boldsymbol{z}_j$ as the residual of the Lasso:

$$\boldsymbol{z}_j = \boldsymbol{x}_j - \boldsymbol{X}_{-j}\widehat{\gamma}_{-j},\ \widehat{\gamma}_{-j} = \arg\min_{\boldsymbol{b}}\Big\{\frac{|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{b}|_2^2}{2n} + \lambda_j|\boldsymbol{b}|_1\Big\}. \tag{6}$$

Explicit choices of λ_j are described in the next subsection. A rationale for the use of a common penalty level λ_j for all components of $\boldsymbol{b}$ in (6) is the normalization of all variables to $|\boldsymbol{x}_k|_2^2 = n$. In an alternative in Subsection 2.3 called restricted LDPE, the penalty is set to zero for certain components of $\boldsymbol{b}$ in (6).

It follows from the Karush-Kuhn-Tucker conditions for (6) that $|\boldsymbol{x}_k^T\boldsymbol{z}_j/n| \le \lambda_j$, so that

$$\Big|\sum_{k\neq j}\boldsymbol{z}_j^T\boldsymbol{x}_k(\beta_k - \widehat{\beta}_k^{(init)})\Big| \le \max_{k\neq j}\big|\boldsymbol{z}_j^T\boldsymbol{x}_k\big|\,|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}|_1 \le n\lambda_j\,|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}|_1. \tag{7}$$

This of course is a conservative bound. Let $\eta_j = n\lambda_j/|\boldsymbol{z}_j|_2$ and $\tau_j = |\boldsymbol{z}_j|_2/|\boldsymbol{z}_j^T\boldsymbol{x}_j|$. Since $\boldsymbol{z}_j^T\boldsymbol{\varepsilon} \sim N(0, \sigma|\boldsymbol{z}_j|_2^2)$, it follows from (5) that

$$\eta_j|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}|_1/\sigma = o(1)\ \Rightarrow\ \tau_j^{-1}\big(\widehat{\beta}_j - \beta_j\big) \approx N(0,\sigma^2). \tag{8}$$

Sufficient conditions for $\eta_j|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}|_1/\sigma = o(1)$ will be given in the next Section.

2.3. Implementation with the scaled Lasso.

We describe specific implementations using scaled Lasso [Ant10, SZ10, SZ11] to provide the initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$, the noise level $\widehat{\sigma}$, and the score vectors $\boldsymbol{z}_j$.

The scaled Lasso is a joint convex minimization method given by

$$\big\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\big\} = \arg\min_{\boldsymbol{b},\sigma}\Big\{\frac{|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}|_2^2}{2\sigma n} + \frac{\sigma}{2} + \lambda_0|\boldsymbol{b}|_1\Big\} \tag{9}$$

with a penalty level λ_0. This automatically provides an estimate of the noise level in addition to the initial estimator of $\boldsymbol{\beta}$. We use $\lambda_0 = \lambda_{univ} = \sqrt{(2/n)\log p}$ in our simulation study. Existing error bounds for the estimation of both $\boldsymbol{\beta}$ and σ require $\lambda_0 = A\sqrt{(2/n)\log(p/\epsilon)}$ with certain $A > 1$ and $0 < \epsilon \le 1$ [SZ11].

The scaled Lasso can be also used to determine λ_j for the $\boldsymbol{z}_j$ in (6). However, the penalty level for the scaled Lasso, set to guarantee performance bounds for the estimation of regression coefficients and noise level, may not be the best for controlling the bias and the standard error of the LDPE. In (7) and (8), we use $\eta_j = \lambda_j/|\boldsymbol{z}_j^T\boldsymbol{x}_j|$ to bound the ratio between the bias and standard error of $\widehat{\beta}_j$ and $\tau_j = |\boldsymbol{z}_j|_2/|\boldsymbol{z}_j^T\boldsymbol{x}_j|$ to approximate the standard error. We choose λ_j by tracking η_j and τ_j in the Lasso path as follows.

The basic idea is to allow some over fitting of $\boldsymbol{x}_j$ as long as τ_j and η_j are reasonably small. Let

$$\begin{aligned}
\widehat{\gamma}_{-j}(\lambda) &= \arg\min_{\boldsymbol{b}}\Big\{|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{b}|_2^2/(2n) + \lambda|\boldsymbol{b}|_1\Big\},\\
\boldsymbol{z}_j(\lambda) &= \boldsymbol{x}_j - \boldsymbol{X}_{-j}\widehat{\gamma}_{-j}(\lambda),\\
\eta_j(\lambda) &= \max_{k\neq j}|\boldsymbol{x}_k^T\boldsymbol{z}_j(\lambda)|/|\boldsymbol{z}_j(\lambda)|_2 = n\lambda/|\boldsymbol{z}_j(\lambda)|_2,\\
\tau_j(\lambda) &= |\boldsymbol{z}_j(\lambda)|_2/|\boldsymbol{x}_j^T\boldsymbol{z}_j(\lambda)|
\end{aligned} \tag{10}$$

be the coefficient estimator $\widehat{\boldsymbol{\gamma}}_{-j}$, residual $\boldsymbol{z}_j$, and factors η_j and τ_j along the Lasso path for regressing $\boldsymbol{x}_j$ against $\boldsymbol{X}_{-j}$. Let $\widehat{\sigma}_j^*$ be scaled Lasso estimate of the noise level in this linear model at a penalty level λ_j^*,

$$\widehat{\sigma}_j^* = \arg\min_{\sigma}\min_{\boldsymbol{b}}\Big\{\frac{|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{b}|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_j^*|\boldsymbol{b}|_1\Big\}. \tag{11}$$

If we used scaled Lasso with $|\boldsymbol{z}_j(\widehat{\sigma}_j^*\lambda_j^*)| = \widehat{\sigma}_j^* n^{1/2}$ to calculate $\widehat{\boldsymbol{\gamma}}_{-j}$, we would pick $\boldsymbol{z}_j(\widehat{\sigma}_j^*\lambda_j^*)$ as $\boldsymbol{z}_j$. This would provide factors $\tau_j(\widehat{\sigma}_j^*\lambda_j^*)$ and $\eta_j(\widehat{\sigma}_j^*\lambda_j^*) = n^{1/2}\lambda_j^* = \sqrt{2\log p}$ in (8).

To increase the accuracy of the coverage probability of the confidence interval for β_j, we reduce the ratio of bias to standard error of our estimator in exchange for a controlled increase in its standard error; specifically, we allow the τ_j to increase by an additional factor κ_0 in order to reduce η_j. Thus, with (10) and (11), we pick

$$\boldsymbol{z}_j = \boldsymbol{z}_j(\lambda_j),\ \lambda_j = \arg\min_{\lambda}\Big\{\eta_j(\lambda) : \tau_j(\lambda) \le (1+\kappa_0)\tau_j(\widehat{\sigma}_j^*\lambda_j^*)\Big\}, \tag{12}$$

where $\kappa_0 > 0$ is a pre-determined constant. This yields $\tau_j = \tau_j(\lambda_j)$ and $\eta_j = \eta_j(\lambda_j)$ in (8).

The following proposition, proved in the appendix, summarizes some useful properties of the procedure (10), (11), and (12).

Proposition 1. *Both functions $|\boldsymbol{z}_j(\lambda)|_2$ and $\eta_j(\lambda)$ are nondecreasing in λ and the function $\tau_j(\lambda)$ is no greater than $1/|\boldsymbol{z}_j(\lambda)|$ in the Lasso path (10). Moreover, it holds in in (12) that*

$$\eta_j(\lambda_j) \le \eta_j(\widehat{\sigma}_j^*\lambda_j^*) = \sqrt{n}\lambda_j^*,\ \tau_j(\lambda_j) \le (1+\kappa_0)/(\widehat{\sigma}_j^* n^{1/2}). \tag{13}$$

Remark 1. *Since $\eta_j(\lambda)$ is a nondecreasing function of λ, (12) can be carried out by minimizing λ under the constraint on $\tau_j(\lambda)$.*

Proposition 1 allows us to set a specific λ_j^* to control η_j and a κ_0 to control the inflation of the standard error from the scaled Lasso. Sensible choices include $\kappa_0 = 1/2$ and $\lambda_j^* = \lambda_{univ} = \sqrt{(2/n)\log p}$ to guarantee $\eta_j \le \sqrt{2\log p}$. Along with (10), (11), and (12), this gives a complete description of a specific implementation. We would like to emphasize here that given κ_0, the score vector $\boldsymbol{z}_j$ in (12) is completely determined by the design matrix $\boldsymbol{X}$.

We have also experimented with an LDPE with a restricted Lasso relaxation for $\boldsymbol{z}_j$. This restricted LDPE can be viewed as a special case of a more general weighted low dimensional projection with different levels of relaxation for different variables $\boldsymbol{x}_k$ according to their correlation to $\boldsymbol{x}_j$. Although we have used (7) to bound the bias, the summands with larger absolute correlation $|\boldsymbol{x}_j^T\boldsymbol{x}_k/n|$ are likely to have a greater contribution to the bias due to initial estimation error $|\widehat{\beta}_k^{(init)} - \beta_k|$. A remedy for this phenomenon is to force smaller $|\boldsymbol{z}_j^T\boldsymbol{x}_k/n|$ for large $|\boldsymbol{x}_j^T\boldsymbol{x}_k/n|$ with a weighted relaxation. For the Lasso (6), this weighted relaxation can be written as

$$\boldsymbol{z}_j = \boldsymbol{x}_j - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}_{-j},\ \widehat{\boldsymbol{\gamma}}_{-j} = \arg\min_{\boldsymbol{b}}\Big\{\frac{|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{b}|_2^2}{2n} + \lambda_j\sum_{k\ne j} w_k|b_k|\Big\}$$

with w_k being a decreasing function of the absolute correlation $|\boldsymbol{x}_j^T\boldsymbol{x}_k/n|$. For the restricted LDPE, we simply set $w_k = 0$ for large $|\boldsymbol{x}_j^T\boldsymbol{x}_k/n|$ and $w_k = 1$ for other k.

Here is a scaled Lasso implementation of this restricted LDPE. Let $K_{j,m}$ be the index set of the m largest $|\boldsymbol{x}_j^T\boldsymbol{x}_k|$ with $k \neq j$ and $\boldsymbol{P}_{j,m}$ be the orthogonal projection to the linear span of $\{\boldsymbol{x}_k, k \in K_{j,m}\}$. We modify the procedure (10), (11), and (12) by first taking the projection of all design vectors to the orthogonal complement of $\{\boldsymbol{x}_k, k \in K_{j,m}\}$:

$$\boldsymbol{z}_j = f(\boldsymbol{P}_{j,m}^{\perp}\boldsymbol{x}_j, \boldsymbol{P}_{j,m}^{\perp}\boldsymbol{X}_{-j}). \tag{14}$$

where $f(\boldsymbol{x}_j, \boldsymbol{X}_{-j})$ denotes the $\boldsymbol{z}_j$ in (12), explicitly as a function of $\boldsymbol{x}_j$ and $\boldsymbol{X}_{-j}$.

2.4. **Confidence intervals.** In Section 3, we will provide sufficient conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ under which the approximation error in (5) is of smaller order than the standard deviation of the noise component. We construct approximate confidence intervals for such configurations of $\{\boldsymbol{X}, \boldsymbol{\beta}\}$ as follows.

The covariance of the noise component in (5) is proportional to

$$\boldsymbol{V} = (V_{jk})_{p\times p},\ V_{jk} = \frac{\boldsymbol{z}_j^T\boldsymbol{z}_k}{|\boldsymbol{z}_j^T\boldsymbol{x}_j||\boldsymbol{z}_k^T\boldsymbol{x}_k|} = \sigma^{-2}\mathrm{Cov}\Big(\frac{\boldsymbol{z}_j^T\boldsymbol{\varepsilon}}{\boldsymbol{z}_j^T\boldsymbol{x}_j}, \frac{\boldsymbol{z}_k^T\boldsymbol{\varepsilon}}{\boldsymbol{z}_k^T\boldsymbol{x}_k}\Big). \tag{15}$$

Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^T$ be the vector of LDPE $\widehat{\beta}_j$ in (4). For example, we may choose $\widehat{\boldsymbol{\beta}}^{(init)}$ in (9) and $\boldsymbol{z}_j$ in (12), (14), or (26) in the construction of $\widehat{\boldsymbol{\beta}}$. For sparse vectors $\boldsymbol{a}$, e.g. $|\boldsymbol{a}|_0 = 2$ for a contrast between two regression coefficients, an approximate $(1-\alpha)100\%$ confidence interval is

$$\big|\boldsymbol{a}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{a}^T\boldsymbol{\beta}\big| \le \widehat{\sigma}\Phi^{-1}(1-\alpha/2)(\boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a})^{1/2}, \tag{16}$$

where $\widehat{\sigma}$ is the scaled Lasso estimator of the noise level σ in (9) and Φ is the standard normal distribution function. An alternative, larger estimate of σ, which produces more conservative approximate confidence intervals, is the penalized maximum likelihood estimator of [SBvdG10].

3. Theoretical Results

We show that when the ℓ_1 loss of the initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$ is of an expected magnitude and the noise level estimator $\widehat{\sigma}$ is consistent, the LDPE based confidence interval has approximately the preassigned coverage probability under a capped ℓ_1 relaxation of the sparsity condition $|\boldsymbol{\beta}|_0 \le s$, provided that $s\log p \ll n^{1/2}$. We review proper conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ under which such convergence of $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$ has already been established. We prove that for certain random design matrices, the width of the confidence interval is of the order $n^{-1/2}$.

3.1. **Deterministic designs.** Here we establish the asymptotic normality of the LDPE (4) and the validity of the resulting confidence interval (16) for deterministic design matrices.

Let $\lambda_{univ} = \sqrt{(2/n)\log p}$. Suppose $\boldsymbol{\beta}$ is sparse in the sense of

$$\textstyle\sum_{j=1}^p \min\big\{|\beta_j|/(\sigma\lambda_{univ}), 1\big\} \le s. \tag{17}$$

This condition holds if $\boldsymbol{\beta}$ is ℓ_0 sparse with $|\boldsymbol{\beta}|_0 \le s$ or ℓ_q sparse with $|\boldsymbol{\beta}|_q^q/(\sigma\lambda_{univ})^q \le s$, $0 < q \le 1$. A generic condition we impose on the initial estimator is

$$P\Big\{|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}|_1 \ge C_1 s\sigma\sqrt{(2/n)\log(p/\epsilon)}\Big\} \le \epsilon \tag{18}$$

for a certain fixed constant C_1 and all $\alpha_0/p^2 \le \epsilon \le 1$, where $\alpha_0 \in (0,1)$ is a preassigned significance level. By requiring a fixed C_1, we implicitly impose regularity conditions on the design $\boldsymbol{X}$ and the sparsity index s in (17). Existing oracle inequalities can be used to verify (18) for various regularized estimators of $\boldsymbol{\beta}$ under different sets of conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ [CT07, ZH08, BRT09, vdGB09, Zha09, Zha10, YZ10, SZ11, ZZ11]. Although most existing results are derived for penalty/threshold levels depending on the noise level σ and under the ℓ_0 sparsity condition, their proofs can be combined or extended to obtain (18). We also impose a similar generic condition on an estimator $\widehat{\sigma}$ for the noise level:

$$P\Big\{|\widehat{\sigma}/\sigma - 1| \ge C_2 s(2/n)\log(p/\epsilon)\Big\} \le \epsilon, \tag{19}$$

with fixed C_2 and all $\alpha_0/p^2 \le \epsilon \le 1$. We use the same ϵ in (18) and (19) without much loss of generality. For the joint estimation of $\{\boldsymbol{\beta}, \sigma\}$ with scaled Lasso (9), a specific set of sufficient conditions for (18) and (19), based on [SZ11], will be stated in Subsection 3.2. In fact, the probability of the union of the two events is smaller than $\epsilon/\max\{1, n^{1/2}\lambda_0\}$ in this specific case with $\lambda_0 = A\sqrt{(2/n)\log(p/\epsilon)}$ for certain $A > 1$.

Theorem 1. *Suppose (1) holds. Let $\widehat{\beta}_j$ be given by (4) with a $\boldsymbol{z}_j$ depending on $\boldsymbol{X}$ only and an initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$. Let $\max(\epsilon'_n, \epsilon''_n) \to 0+$, $\tau_j = |\boldsymbol{z}_j|_2/|\boldsymbol{x}_j^T\boldsymbol{z}_j|$, and $\eta_j = \max_{k\ne j}|\boldsymbol{x}_k^T\boldsymbol{z}_j|/|\boldsymbol{z}_j|_2$. Suppose (18) holds and $\eta_j C_1 s\sigma\sqrt{(2/n)\log(p/\epsilon)} \le \epsilon'_n$ for all j. Then,*

$$\max_j P\Big\{\big|\tau_j^{-1}(\widehat{\beta}_j - \beta_j) - \boldsymbol{z}_j^T\boldsymbol{\varepsilon}/|\boldsymbol{z}_j|_2\big| > \epsilon'_n\Big\} \le \epsilon. \tag{20}$$

If in addition (19) holds with $C_2 s(2/n)\log(p/\epsilon) \le \epsilon''_n$, then for all $j \le p$ and $t \in \mathbb{R}$,

$$\Phi(t - \epsilon'_n - \epsilon''_n|t|) - 2\epsilon \le P\Big\{\tau_j^{-1}(\widehat{\beta}_j - \beta_j) \le \widehat{\sigma}t\Big\} \le \Phi(t + \epsilon' + \epsilon''_n|t|) + 2\epsilon. \tag{21}$$

Consequently, for the covariance matrix $\boldsymbol{V}$ in (15) and all fixed m,

$$\lim_{n\to\infty}\inf_{|\boldsymbol{a}|_0 \le m} P\Big\{\big|\boldsymbol{a}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{a}^T\boldsymbol{\beta}\big| \le \widehat{\sigma}\Phi^{-1}(1-\alpha/2)(\boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a})^{1/2}\Big\} = 1 - \alpha. \tag{22}$$

Remark 2. *In our implementation (12) with $\lambda_j^* = \lambda_{univ}$, $\boldsymbol{z}_j$ is the residual of the Lasso estimator in the regression model for $\boldsymbol{x}_j$ against $\boldsymbol{X}_{-j} = (\boldsymbol{x}_k, k \ne j)$ with a penalty level λ_j to guarantee $\eta_j \le \sqrt{2\log p}$. Thus, the dimension constraint for the asymptotic normality and proper coverage probability in Theorem 1 is $s(\log p)/\sqrt{n} \to 0$. It is expected from existing theoretical results for the Lasso and Proposition 1 that $\tau_j \le 1/|\boldsymbol{z}_j|_2 \asymp n^{-1/2}$ in (12).*

Since $(\boldsymbol{z}_j^T\boldsymbol{\varepsilon}/|\boldsymbol{z}_j|_2, j \le p)$ has a multivariate normal distribution with identical marginal distributions $N(0,\sigma^2)$, (20) establishes the joint asymptotic normality of the LDPE for finitely many $\widehat{\beta}_j$ under (18). Under the additional condition (19), (21) and (22) justify the approximate coverage probability of the resulting confidence intervals. Sufficient conditions for (18) and (19) are given in Subsection 3.2, and the convergence rate $\tau_j \le 1/|\boldsymbol{z}_j|_2 \asymp n^{-1/2}$ is verified in Subsection 3.3.

3.2. **Oracle inequalities.** We focus here on the scaled Lasso (9) as a specific choice of the initial estimator, since the confidence interval in Theorem 1 is based on the joint estimation of regression coefficients and noise level.

Let $\xi \geq 1$ and $S = \{j : |\beta_j| > \sigma\lambda_{univ}\}$. Define a sign-restricted cone invertibility factor

$$\text{(23)}\quad \mathrm{SCIF}(\xi, S) = \inf\left\{|\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u}|_\infty|S|/(n|\boldsymbol{u}_S|_1) : |\boldsymbol{u}_{S^c}|_1 \leq \xi|\boldsymbol{u}_S|_1, u_j\boldsymbol{x}_j^T\boldsymbol{X}\boldsymbol{u} \leq 0, j \notin S\right\}$$

as in [YZ10]. Since $|\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u}|_\infty|S|/(n|\boldsymbol{u}_S|_1) \geq |\boldsymbol{X}\boldsymbol{u}|^2|S|/(n|\boldsymbol{u}_S|_1^2)$ for vectors $\boldsymbol{u}$ satisfying the restrictions in (23), $\mathrm{SCIF}(\xi, S)$ is a slightly larger quantity than the compatibility factor [vdGB09]. The following theorem is a consequence of Theorem 2 in [SZ11].

Theorem 2. *Let $\{A, \xi, c_0\}$ be fixed positive constants with $\xi > 1$ and $A > (\xi+1)/(\xi-1)$. Let $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$ be as in (9) with $\lambda_0 = A\sqrt{(2/n)\log(p/\epsilon)}$. Suppose $\min_{|S|\leq s} SCIF(\xi, S) \geq c_0$ and (17) holds with $s(\log p)/n \to 0$ as $n \to \infty$. Then, for sufficiently large n, (18) and (19) hold with C_1 and C_2 depending on $\{A, \xi, c_0\}$ only. Consequently, (20), (21), and (22) hold.*

The main condition of Theorem 2 is $\min_{|S|\leq s} \mathrm{SCIF}(\xi, S) \geq c_0$ for $s(\log p)/n \to 0$. This condition holds if the sparse Riesz condition holds: $c_* \leq$eigenvalue$(\boldsymbol{X}_S^T\boldsymbol{X}_S/n) \leq c^*$ for all $|S| \leq (c^*/c_* + 1/2)s$ [ZH08, YZ10, Zha10]. This main condition of Theorem 2 holds for a certain class of random design matrices $\boldsymbol{X}$ described in Subsection 3.3.

3.3. **Random designs.** We assume in this subsection that $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is a column normalized version of a Gaussian random matrix $\widetilde{\boldsymbol{X}}$,

$$\text{(24)}\qquad \boldsymbol{x}_j = \widetilde{\boldsymbol{x}}_j\sqrt{n}/|\widetilde{\boldsymbol{x}}_j|_2,\ \ \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)\ \text{ has iid } N(0, \boldsymbol{\Sigma}) \text{ rows.}$$

We assume without loss of generality that the diagonal elements of $\boldsymbol{\Sigma}$ all equal to 1. In this setting, $\boldsymbol{x}_j$ is related to other design variables $\boldsymbol{X}_{-j}$ through a linear model

$$\text{(25)}\qquad \boldsymbol{x}_j = \boldsymbol{X}_{-j}\boldsymbol{\gamma}_{-j} + \boldsymbol{e}^{(j)},\ \ |\widetilde{\boldsymbol{x}}_j|_2 n^{-1/2}\boldsymbol{e}^{(j)} \sim N(0, \sigma_j^2 I_{n\times n}).$$

This model is a motive for the use of the Lasso in (10), (11), and (12). However, the goal of the procedure is to find $\boldsymbol{z}_j$ with suitable τ_j and η_j for controlling the variance and bias of the LDPE (4) as in Theorem 1. The following theorem justifies low-dimensional statistical inference based on LDPE (4) for the random design (24) by verifying all conditions of Theorems 1 and 2 under the sparsity condition (17) with $s(\log p)/\sqrt{n} \to 0$. Define a class of vectors with small ℓ_q tail as

$$\mathscr{B}_q(s, \lambda) = \left\{\boldsymbol{b} \in \mathbb{R}^p : \textstyle\sum_{j=1}^p \min(|b_j|^q/\lambda^q, 1) \leq s\right\}.$$

Theorem 3. *Suppose (1) and (24) hold with $diag(\boldsymbol{\Sigma}) = \boldsymbol{I}_{p\times p}$ and $eigenvalues(\boldsymbol{\Sigma}) \subset [c_*, c^*]$, where $0 < c_* < c^* < \infty$ are fixed. Let $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$ be as in (9) with $\lambda_0 = A\sqrt{(2/n)\log(p/\epsilon)}$ for a certain fixed $A > 1$. Let $\boldsymbol{z}_j$ be as in (12) with $\lambda_j^* \asymp \lambda_{univ}$ and $\widehat{\beta}_j$ be as in (4). Suppose (17) holds with $s(\log p)/n^{1/2} \to 0$. Then, $\min_{|S|\leq s} SCIF(\xi, S) \geq c_0$ in (23) with fixed $c_0 > 0$. Consequently, (20), (21) and (22) hold as in Theorem 2. Moreover, if $s^*(\log p)/n \to 0$, then $\tau_j \lesssim n^{-1/2}$ whenever $\boldsymbol{\gamma}_{-j} \in \mathscr{B}_2(s^*, \lambda_0)$.*

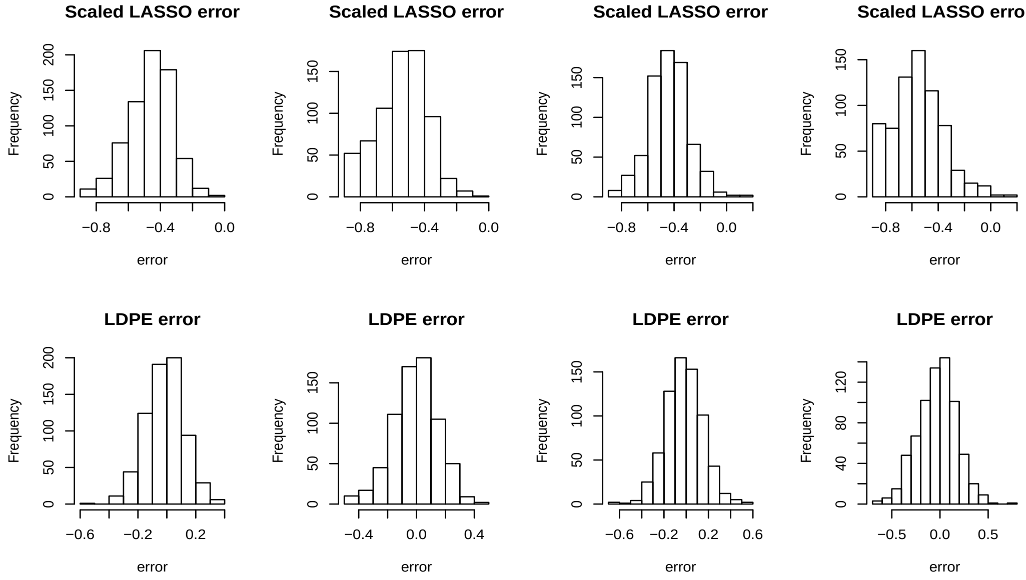

Figure 1. Histogram of errors when estimating maximal β_j using the scaled Lasso and LDPE, in simulation settings (A), (B), (C), and (D), from left to right.

Remark 3. *It follows from the block matrix inversion formula that in (25)*

$$\boldsymbol{\gamma}_{-j} = \Big((\boldsymbol{\Sigma}^{-1})_{jk}|\widetilde{\boldsymbol{x}}_k|_2/|\widetilde{\boldsymbol{x}}_j|_2, k \neq j\Big),\ \ \sigma_j^2 = (\boldsymbol{\Sigma}^{-1})_{jj}.$$

Since $|\widetilde{\boldsymbol{x}}_j|_2^2 \sim \chi_n^2$, the condition $\boldsymbol{\gamma}_{-j} \in \mathscr{B}_2(s^*, \lambda_0)$ in Theorem 3 can be replaced by $((\boldsymbol{\Sigma}^{-1})_{jk}, k \neq j) \in \mathscr{B}_2(s^*, \lambda_0)$. This does not add much restriction to the condition that eigenvalues$(\boldsymbol{\Sigma}) \subset [c_*, c^*]$.

4. Simulation Results

Our simulation design is set to clearly violate the assumptions of theorems in Section 3. We set $n = 200$, $p = 3000$, $\lambda_j^* = \lambda_{univ} = \sqrt{(2/n)\log p}$, $\beta_j = 3\lambda_{univ}$ for $j = 1500, 1800, 2100, \ldots, 3000$, and $\beta_j = 3\lambda_{univ}/j^\alpha$ for all other j. This gives $(s, s{*}(\log p)/n^{1/2}) = (8.93, 5.05)$ and $(29.24, 16.55)$ respectively for $\alpha = 2$ and 1, while the theorems require $s(\log p)/\sqrt{n} \to 0$, where $s = \sum_j \min(|\beta_j|/\lambda_{univ}, 1)$. We run simulation experiments with 100 replications in each setting. In each replication, we generate an independent copy of $(\widetilde{\boldsymbol{X}}, \boldsymbol{X}, \boldsymbol{y})$, where $\widetilde{X} = (\widetilde{x}_{ij})_{n\times p}$ has iid $N(0, \Sigma)$ rows with $\Sigma = (\rho^{|j-k|})_{p\times p}$, $\boldsymbol{x}_j = \widetilde{\boldsymbol{x}}_j\sqrt{n}/|\widetilde{\boldsymbol{x}}_j|_2$, and $(\boldsymbol{X}, \boldsymbol{y})$ as in (1) with $\sigma = 1$. This example includes four cases, labeled (A), (B), (C), and (D), respectively: $(\alpha, \rho) = (2, 1/5)$, $(1, 1/5)$, $(2, 4/5)$, and $(1, 4/5)$, with case (D) being the most difficult one.

In addition to the Lasso with penalty level $\lambda = \lambda_{univ}$ and the scaled Lasso (9), we consider the LDPE and its restricted version with (14). For both the LDPE and restricted LDPE,

		Estimator		
		Lasso	scaled Lasso	LDPE
(A)	mean	-0.2946	-0.4601	-0.0149
	variance	0.0090	0.0187	0.0175
	median	0.2941	0.4459	0.0885
(B)	mean	-0.2998	-0.5392	-0.0075
	variance	0.0108	0.0239	0.0239
	median	0.2949	0.5296	0.0993
(C)	mean	-0.3009	-0.4406	-0.0184
	variance	0.0135	0.0229	0.0280
	median	0.3054	0.4411	0.1088
(D)	mean	-0.3214	-0.5514	-0.0290
	variance	0.0193	0.0351	0.0406
	median	0.3179	0.5581	0.1315

Table 1. Summary statistics for the errors of the Lasso with fixed penalty $\lambda = \lambda_{univ}$, the scaled Lasso, and LDPE estimates of the seven maximal $\beta_j = |\boldsymbol{\beta}|_\infty$.

the scaled Lasso is used to generate $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$ and the procedure (11), (10), and (12) is used to generate $\boldsymbol{z}_j$, with $\kappa_0 = 1/2$ and $\lambda_j^* = \lambda_{univ}$ to guarantee $\eta_j \le \sqrt{2\log p}$. For the restricted LDPE, $m = 20$ is used in (14)

The asymptotic normality of the LDPE holds well in our simulation experiments. Table 1 and Figure 1 demonstrate the behavior of the LDPE for the largest β_j, compared with the Lasso. For a small increase in variance, the LDPE significantly reduces the bias of the Lasso and scaled Lasso estimates. The scaled Lasso has more bias than the Lasso, but is entirely data-driven. These results hold over all simulation settings. The LDPE shows the largest improvement in performance when estimating large β_j. Although the asymptotic normality of the LDPE holds even better for small β_j in the simulation study, a parallel comparison for small β_j is not meaningful; the Lasso typically estimates small β_j by zero, while the raw LDPE is not designed to be sparse.

	(A)	(B)	(C)	(D)
coverage	0.9578	0.9602	0.9502	0.9501

Table 2. Mean coverage for LDPEs for all β_j

The overall coverage probability of the LDPE-based confidence interval matches well to the preassigned level, as expected from the asymptotic normality. The LDPE creates confidence intervals $\widehat{\beta}_j \pm 1.96\widehat{\sigma}\tau_j$ with approximately 95% coverage. Refer to Table 2 for precise

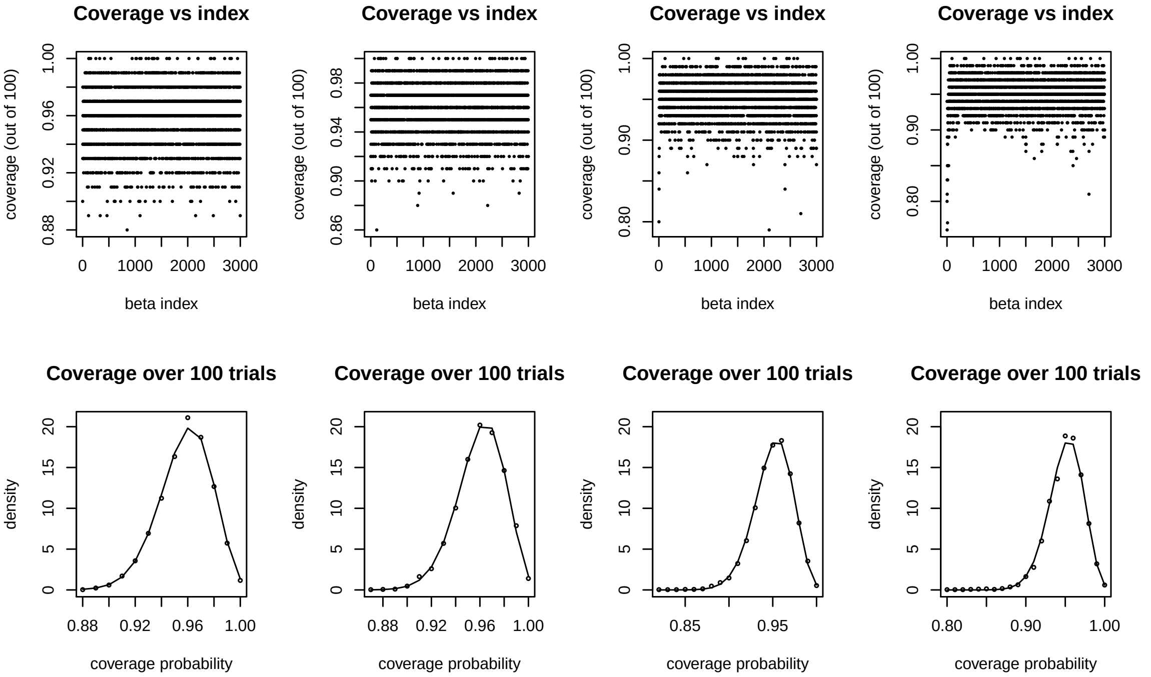

FIGURE 2. Top: Relative coverage frequencies versus the index of β_j. Bottom: The percentage of variables for given values of the relative coverage frequency, superimposed with the binomial$(100, 0.95)$ probability mass function. Figures depict results from simulations (A), (B), (C), and (D), from left to right.

values. Moreover, the empirical distribution of the simulated relative coverage frequencies matches that of the binomial$(100, 0.95)$ very well in all four settings, as shown in the bottom row of Figure 2.

Although the coverage probability is satisfactory over all β_j, we have some degree of under-coverage when large values of β_j are associated with highly correlated columns of $\boldsymbol{X}$. This is most apparent when plotting coverage vs. index in simulation (D), but is also visible in simulation (C). In the top row of Figure 2, the range of simulated relative coverage frequencies in settings (C) and (D) is considerably larger, due to a few instances of low coverage for the smallest indices. Note that these are the two simulations with higher correlation between adjacent columns of $\boldsymbol{X}$, and the first few β_j are all relatively large.

The restricted LDPE (14) eliminates the bias caused by relatively large values of β_j associated with highly correlated columns of $\boldsymbol{X}$. Figure 3 shows that this reduces the bias originating from other large values of β and improve coverage probabilities, at the cost of an increase in the variance of the estimator.

We may also consider the performance of LDPE as a point estimator. Figure 4 shows the behavior of the median absolute error for the estimation of β_j. The Lasso and scaled Lasso

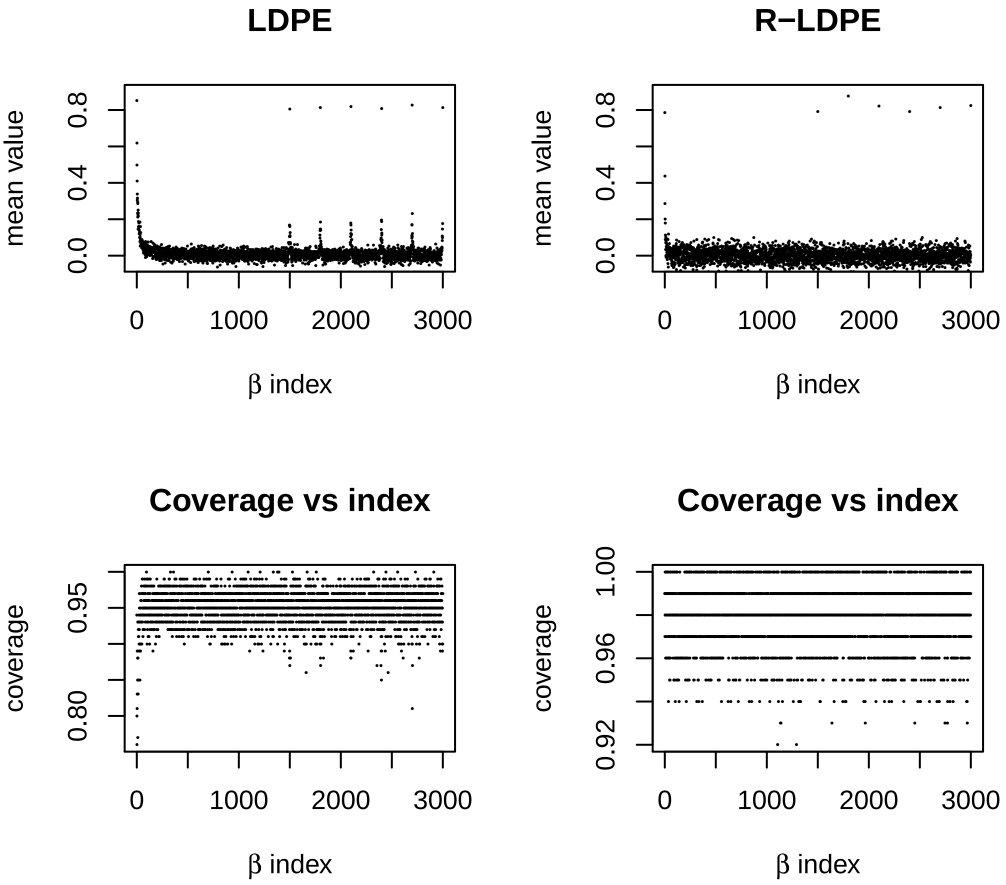

FIGURE 3. Plots of the median of $\widehat{\beta}_j$ (top) and relative coverage frequencies (bottom) over 100 replications for the LDPE (left) and the restricted LDPE (right). The restricted LDPE has smaller bias and more accurate coverage probabilities, but larger variability.

estimators have large biases for bigger values of β_j but perform very well for smaller values. On the other hand, the median absolute error for the LDPE is very stable since it is not designed to be sparse without post processing. They perform significantly better than the scaled Lasso (the initial estimator of the LDPE) for large β_j, but Lasso outperforms LDPE as a point estimator for smaller β_j. This can be adjusted by thresholding the LDPE.

Once we have thresholded, we may consider the L^2 loss of the LDPE as an estimate of the vector $\boldsymbol{\beta}$. The mean, standard deviation, and median L^2 loss over 100 replications is listed in Table 3. In the simplest case, (A), the thresholded LDPE outperforms both the Lasso and the scaled Lasso in terms of L^2 loss. The mean and median L^2 loss for the Lasso and LDPE is comparable in simulations (B) and (C), with both outperforming the scaled

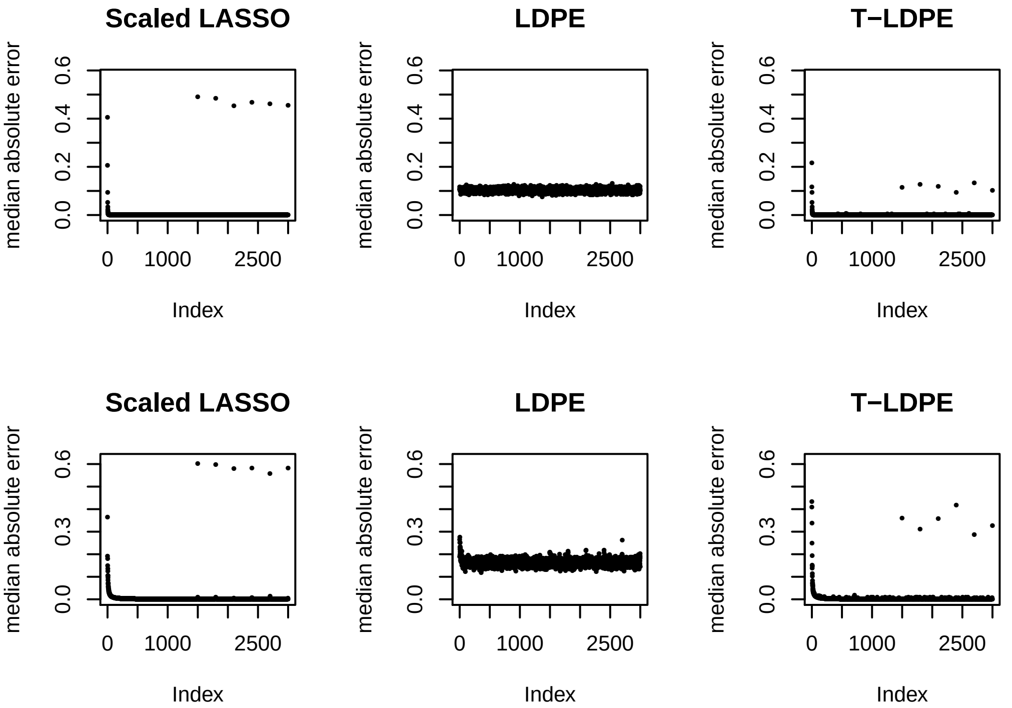

FIGURE 4. Plots of median absolute error for scaled Lasso and LDP estimators, in simulations (A) and (D). The patterns in simulations (B) and (C) are similar to those in (A) and (D), respectively.

Lasso. However, the standard deviation of the L^2 loss is larger for the LDPE. In the hardest case, (D), which has both a high correlation between adjacent columns of $\boldsymbol{X}$ and a slower decay for $\boldsymbol{\beta}$, the performance of the LDPE is similar to that of the scaled Lasso, with the Lasso with fixed penalty $\lambda = \sigma\lambda_{univ}$ outperforming both data-driven estimators.

5. Discussion

We have developed the LDPE method of constructing $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ for the individual regression coefficients and estimators for their finite dimensional covariance structure. Under proper conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$, we have proven the asymptotic unbiasedness and normality of the finite dimensional distribution functions of these estimators and the consistency of their estimated covariances. This allows one to assess the level of significance of each unknown coefficient β_j more accurately than commonly used estimators of the entire vector $\boldsymbol{\beta}$. The raw LDPE estimator is not sparse, but it can be thresholded to take advantage of the sparsity of $\boldsymbol{\beta}$. Compared with existing variable selection methods, thresholding the LDPE allows selection of large regression coefficients in the presence of many small regression coefficients, and the sampling distribution of the thresholded LDPE can be bounded

		Estimator		
		Lasso	scaled Lasso	T-LDPE
(A)	mean	0.7235	1.6693	0.3072
	sd	0.1965	0.6738	0.3599
	median	0.6918	1.4946	0.1741
(B)	mean	1.0023	2.6109	1.0233
	sd	0.2286	0.8202	0.7669
	median	0.9584	2.4635	0.6619
(C)	mean	0.7796	1.5717	0.9990
	sd	0.2250	0.5952	0.9033
	median	0.7390	1.4266	0.8426
(D)	mean	1.1555	2.6811	2.9132
	sd	0.3056	0.7694	1.3340
	median	1.1349	2.7298	2.8206

TABLE 3. Summary statistics of L^2 loss for the Lasso with $\lambda = \sigma\lambda_{univ}$, scaled Lasso, and thresholded LDPE estimates of $\boldsymbol{\beta}$

based on our theoretical results. For example, we allow ϵ to range in $[\alpha_0/p^2, 1]$ in Section 3 to cover further calculations involving Bonferroni and other multiplicity adjustments.

In this paper, we use the Lasso to provide a relaxation of the projection of $\boldsymbol{x}_j$ to $\boldsymbol{x}_j^{\perp}$. This choice is primarily due to our familiarity with the computation of the Lasso and the readily available scaled Lasso method of choosing a penalty level. We have also considered some other methods of relaxing the projection. Among these other methods, a particularly interesting one is the following constrained minimization of the variance of the noise term in (5):

$$\boldsymbol{z}_j = \arg\min_{\boldsymbol{z}} \Big\{ |\boldsymbol{z}|_2^2 : |\boldsymbol{z}_j^T \boldsymbol{x}_j| = n, \max_{k \neq j} |\boldsymbol{z}_j^T \boldsymbol{x}_k / n| \le \lambda_j' \Big\}. \tag{26}$$

Similar to the Lasso in (6), (26) is quadratic programming. The Lasso solution (6) is feasible in (26) with $\lambda_j n/|\boldsymbol{z}_j^T \boldsymbol{x}_j| = \lambda_j'$. Our results on these and other extensions of our ideas and methods will be presented in a forthcoming paper.

6. Appendix

Proof of Proposition 1. Since $\widehat{\boldsymbol{\gamma}}_{-j}(\lambda)$ is continuous and piecewise linear in λ, it suffices to consider a fixed open interval $\lambda \in I_0$ in which $A = \{k \neq j : w_k(\lambda) \neq 0\}$ and $\boldsymbol{s} = \mathrm{sgn}(\widehat{\boldsymbol{\gamma}}_{-j}(\lambda))$ do not change with λ. It follows from the Karush-Kuhn-Tucker conditions for the Lasso that

$$\boldsymbol{X}_A^T \{\boldsymbol{x}_j - \boldsymbol{X}_A \widehat{\boldsymbol{\gamma}}_A(\lambda)\}/n = \boldsymbol{X}_A^T \{\boldsymbol{x}_j - \boldsymbol{X}_{-j} \widehat{\boldsymbol{\gamma}}_{-j}(\lambda)\}/n = \lambda \boldsymbol{s}_A.$$

This gives $(\partial/\partial\lambda)\widehat{\boldsymbol{\gamma}}_A(\lambda) = -(\boldsymbol{X}_A^T\boldsymbol{X}_A/n)^{-1}\boldsymbol{s}_A$ for all $\lambda \in I_0$. It follows that

$$\begin{aligned}(\partial/\partial\lambda)|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}_{-j}(\lambda)|_2^2 &= -2\{(\partial/\partial\lambda)\widehat{\boldsymbol{\gamma}}_A(\lambda)\}^T\boldsymbol{X}_A^T(\boldsymbol{x}_j - \boldsymbol{X}_A\widehat{\boldsymbol{\gamma}}_A(\lambda))\\ &= 2\{(\boldsymbol{X}_A^T\boldsymbol{X}_A/n)^{-1}\boldsymbol{s}_A\}^T\boldsymbol{X}_A^T(\boldsymbol{x}_j - \boldsymbol{X}_A\widehat{\boldsymbol{\gamma}}_A(\lambda))\\ &= (2/\lambda)|\boldsymbol{P}_A(\boldsymbol{x}_j - \boldsymbol{X}_A\widehat{\boldsymbol{\gamma}}_A(\lambda))|_2^2,\end{aligned}$$

where $\boldsymbol{P}_A = \boldsymbol{X}_A(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^T$ is the projection to the column space of $\boldsymbol{X}_A$. Thus, $|\boldsymbol{z}_j(\lambda)|_2$ is nondecreasing in λ. Moreover,

$$\begin{aligned}(\lambda^3/2)(\partial/\partial\lambda)\big\{\eta(\lambda)/n\big\}^{-2} &= (\lambda^3/2)(\partial/\partial\lambda)\big\{\lambda^{-2}|\boldsymbol{x}_j - \boldsymbol{X}_A\widehat{\boldsymbol{\gamma}}_A(\lambda)|_2^2\big\}\\ &= |\boldsymbol{P}_A\{\boldsymbol{x}_j - \boldsymbol{X}_A\widehat{\boldsymbol{\gamma}}_A(\lambda)\}|_2^2 - |\boldsymbol{x}_j - \boldsymbol{X}_A\widehat{\boldsymbol{\gamma}}_A(\lambda)|_2^2 \le 0,\end{aligned}$$

so that $\eta(\lambda)$ is nondecreasing in λ. Since $\boldsymbol{x}_j^T\boldsymbol{z}_j(\lambda) = |\boldsymbol{z}_j(\lambda)|_2^2 + \{\boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}_{-j}(\lambda)\}^T\boldsymbol{z}_j(\lambda) = |\boldsymbol{z}_j(\lambda)|_2^2 + \lambda|\widehat{\boldsymbol{\gamma}}_{-j}(\lambda)|_1$, we also have $\tau(\lambda) \le 1/\boldsymbol{z}_j(\lambda)$. Finally, (13) follows since $\lambda_j \le \sigma_j^*\lambda_j^*$ and $|\boldsymbol{z}_j(\widehat{\sigma}_j^*)|_2 = \widehat{\sigma}_j^* n^{1/2}$. □

Proof of Theorem 1. Since $\boldsymbol{z}_j$ is a determined by $\boldsymbol{X}$, (5) and (7) imply

$$\Big|\tau_j^{-1}(\widehat{\beta}_j - \beta_j) - \boldsymbol{z}_j^T\boldsymbol{\varepsilon}/|\boldsymbol{z}_j|_2\Big| \le \Big(\max_{k\neq j}|\boldsymbol{z}_j^T\boldsymbol{x}_k|/|\boldsymbol{z}_j|_2\Big)|\widehat{\boldsymbol{\beta}}^{(init)} - \widehat{\boldsymbol{\beta}}|_1 = \eta_j|\widehat{\boldsymbol{\beta}}^{(init)} - \widehat{\boldsymbol{\beta}}|_1.$$

This and (18) yield (20) as well as the validity of $\boldsymbol{V}$ in (15) as the approximate covariance between $\widehat{\beta}_j$ and $\widehat{\beta}_k$. The rest of the Theorem then follows directly from (19). □

Proof of Theorem 2. Let $\sigma^* = |\boldsymbol{\varepsilon}|_2/\sqrt{n}$. It follows from Theorem 2 in [SZ11] that

$$P\Big\{|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \le C_0\sigma^* s\lambda_0, |\widehat{\sigma}/\sigma^* - 1| \le C_0 s\lambda_0^2\Big\} \ge 1 - \epsilon,$$

with a constant C_0 depending on $\{A, \xi, c_0\}$ only. Since $n(\sigma^*/\sigma)^2$ has the χ_n^2 distribution, (18) and (19) hold with C_1 and C_2 slightly different from C_0. □

Proof of Theorem 3. Since the eigenvalues of $\boldsymbol{\Sigma}$ are uniformly bounded, $\min_{|S|\le s}\mathrm{SCIF}(\xi, S)$ is uniformly bounded away from zero for small $|S|(\log p)/n$ [ZH08, YZ10]. This implies the condition of Theorem 2.

Consider a fixed $j \in J_n$. Let $\widehat{\sigma}_j$ be the scaled Lasso estimator of σ_j in (25) with penalty level λ_0 and S_j be the index set of the elements of $\boldsymbol{\gamma}_{-j}$ with absolute value $\sigma_j\lambda_0$ or larger. Since $E|\boldsymbol{X}_{S_j^c}\boldsymbol{\gamma}_{S_j^c}|_2^2/n \lesssim O(c^*)|\boldsymbol{\gamma}_{S_j^c}|_2^2 \lesssim O(s^*\lambda_0^2) = o(1)$ and the SCIF is uniformly bounded away from zero, Theorem 1 of [SZ11] implies $\widehat{\sigma}_j/\sigma_j = 1 + o(1)$. Thus, in the Lasso path (10), $\tau_j(\widehat{\sigma}_j\lambda_0) \le 1/|\boldsymbol{z}_j(\widehat{\sigma}_j\lambda_0)|_2 = 1/(\widehat{\sigma}_j n^{1/2}) \asymp n^{-1/2}$.

Now consider (12) with penalty level $\lambda_j^* \asymp \lambda_0$. By Proposition 1, both $|\boldsymbol{z}_j(\lambda)|_2$ and $\eta_j(\lambda) = n\lambda/|\boldsymbol{z}_j(\lambda)|_2$ are nondecreasing in λ, so that $\widehat{\sigma}_j \asymp \widehat{\sigma}_j^*$ in (11). Consequently, $\tau_j(\lambda_j) \lesssim n^{-1/2}$ by Proposition 1. □

References

[Ant10] A. Antoniadis, *Comments on: ℓ_1-penalization for mixture regression models*, Test **19** (2010), no. 2, 257–258.

[BRT09] Peter Bickel, Yaacov Ritov, and Alexandre Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Annals of Statistics **37** (2009), no. 4, 1705–1732.

[BvdG11] Peter Bühlmann and Sara van de Geer, *Statistics for high-dimensional data: Methods, theory and applications*, Springer, New York, 2011.

[CDS01] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Review **43** (2001), 129–159.

[CT07] E. Candes and T. Tao, *The dantzig selector: statistical estimation when p is much larger than n (with discussion)*, Annals of Statistics **35** (2007), 2313–2404.

[FF93] I.E. Frank and J.H. Friedman, *A statistical view of some chemometrics regression tools (with discussion)*, Technometrics **35** (1993), 109–148.

[FL01] Jianqing Fan and Runze Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association **96** (2001), 1348–1360.

[FL08] Jianqing Fan and Jinchi Lv, *Sure independence screening for ultrahigh dimensional feature space (with discussion)*, J. R. Statist. Soc. **B, 70** (2008), 849–911.

[FL10] ———, *A selective overview of variable selection in high dimensional feature space*, Statistica Sinica **20** (2010), 101–148.

[FP04] J. Fan and H. Peng, *On non-concave penalized likelihood with diverging number of parameters*, Annals of Statistics **32** (2004), 928–961.

[GR04] E. Greenshtein and Y. Ritov, *Persistence in high–dimensional linear predictor selection and the virtue of overparametrization*, Bernoulli **10** (2004), 971–988.

[Gre06] E. Greenshtein, *Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint*, Annals of Statistics **34** (2006), 2367–2386.

[HMZ08] J. Huang, S. Ma, and C.-H. Zhang, *Adaptive lasso for sparse high-dimensional regression models*, Statistica Sinica **18** (2008), 1603–1618.

[KCO08] Yongdai Kim, Hosik Choi, and Hee-Seok Oh, *Smoothly clipped absolute deviation on high dimensions*, Journal of American Statistical Association **103** (2008), 1665–1673.

[KLT11] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, *Nuclear norm penalization and optimal rates for noisy low rank matrix completion*, Annals of Statistics (2011), to appear.

[Kol09] V. Koltchinskii, *The dantzig selector and sparsity oracle inequalities*, Bernoulli **15** (2009), 799–828.

[MB06] Nicolai Meinshausen and Peter Bühlmann, *High-dimensional graphs and variable selection with the lasso*, Annals of Statistics **34** (2006), 1436–1462.

[MB10] ———, *Stability selection (with discussion)*, Journal of the Royal Statistical Society, B **72** (2010), 417–473.

[MY09] N. Meinshausen and B. Yu, *Lasso-type recovery of sparse representations for high-dimensional data*, Annals of Statistics **37** (2009), 246–270.

[SBvdG10] N. Städler, P. Bühlmann, and S. van de Geer, *ℓ_1-penalization for mixture regression models (with discussion)*, Test **19** (2010), no. 2, 209–285.

[SZ10] Tingni Sun and Cun-Hui Zhang, *Comments on: ℓ_1-penalization for mixture regression models*, Test **19** (2010), no. 2, 270–275.

[SZ11] Tungni Sun and Cun-Hui Zhang, *Scaled sparse linear regression*, Tech. Report arXiv:1104.4595, arXiv, 2011.

[Tib96] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **58** (1996), 267–288.

[Tro06] J. A. Tropp, *Just relax: convex programming methods for identifying sparse signals in noise*, IEEE Transactions on Information Theory **52** (2006), 1030–1051.

[vdGB09] S. van de Geer and P. Bühlmann, *On the conditions used to prove oracle results for the lasso*, Electronic Journal of Statistics **3** (2009), 1360–1392.

[Wai09] M. J. Wainwright, *Sharp thresholds for noisy and high–dimensional recovery of sparsity using ℓ_1–constrained quadratic programming (lasso)*, IEEE Transactions on Information Theory **55** (2009), 2183–2202.

[YZ10] Fei Ye and Cun-Hui Zhang, *Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls*, Journal of Machine Learning Research **11** (2010), 3481–3502.

[ZH08] Cun-Hui Zhang and Jian Huang, *The sparsity and bias of the Lasso selection in high-dimensional linear regression*, Annals of Statistics **36** (2008), no. 4, 1567–1594.

[Zha09] Tong Zhang, *Some sharp performance bounds for least squares regression with L_1 regularization*, Ann. Statist. **37** (2009), no. 5A, 2109–2144.

[Zha10] Cun-Hui Zhang, *Nearly unbiased variable selection under minimax concave penalty*, The Annals of Statistics **38** (2010), 894–942.

[Zha11a] Tong Zhang, *Adaptive forward-backward greedy algorithm for learning sparse representations*, IEEE Transactions on Information Theory (2011), to appear.

[Zha11b] ———, *Multi-stage convex relaxation for feature selection*, Tech. Report arXiv:1106.0565, arXiv, 2011.

[ZL08] Hui Zou and Runze Li, *One-step sparse estimates in nonconcave penalized likelihood models*, Annals of Statistics **36** (2008), no. 4, 1509–1533.

[Zou06] Hui Zou, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association **101** (2006), 1418–1429.

[ZY06] Peng Zhao and Bin Yu, *On model selection consistency of Lasso*, Journal of Machine Learning Research **7** (2006), 2541–2567.

[ZZ11] Cun-Hui Zhang and Tong Zhang, *A general theory of concave regularization for high dimensional sparse estimation problems*, Tech. Report arXiv:1108.4988, arXiv, 2011.

Department of Statistics and Biostatistics, Hill Center, Busch Campus, Rutgers University, Piscataway, NJ 08854, USA
E-mail address: `czhang@stat.rutgers.edu`

Department of Statistics, Columbia University, New York, NY 10027
E-mail address: `sszhang@stat.columbia.edu`